Cláudio Ratke

ALGORITMO PARA IDENTIFICAÇÃO DE CARACTERISTICAS PARA AMOSTRAGEM ESTRATIFICADA

Florianópolis – SC
2006

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Cláudio Ratke

ALGORITMO PARA IDENTIFICAÇÃO DE CARACTERISTICAS PARA AMOSTRAGEM ESTRATIFICADA

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. Dalton Francisco de Andrade
Orientador

Florianópolis, Julho de 2006

ALGORITMO PARA IDENTIFICAÇÃO DE CARACTERISTICAS PARA A ESTRATIFICAÇÃO

Cláudio Ratke

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, Área de Concentração Análise de Dados e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____
Prof. Raul S. Wazlawick, Dr.
Coordenador

Banca Examinadora:

_____
Prof. Dalton Francisco de Andrade, Dr.
Orientador

_____
Prof. Wilton de Oliveira Bussab, Dr.

_____
Prof. Paulo José Ogliari, Dr.

_____
Prof. Pedro Alberto Barbetta, Dr.

_____
Prof. Oscar Dalfovo, Dr.

Dedico este trabalho aos meus pais, Herdi e Wally (in memorium), e a minha esposa Isabel, pelo incentivo recebido durante o desenvolvimento do mesmo.

AGRADECIMENTOS

Agradeço a Deus a oportunidade de fazer o mestrado.

Agradeço ao Prof. Dalton de Andrade pelo apoio, paciência, amizade e orientação.

Agradeço a Universidade Federal de Santa Catarina (UFSC), especialmente ao Departamento de Informática e Estatística (INE), que proporcionou um ótimo ambiente para o meu desenvolvimento profissional e humano.

Agradeço aos prof. Pedro Alberto Barbetta e Paulo José Ogliari que pelas discussões técnicas que enriqueceram o trabalho.

Ao grande amigo e colega de estudos Gilvan Justino, que colaborou de forma decisiva para a realização deste trabalho, desde a definição do tema até sua conclusão.

Aos supermercados Hippo, na pessoa do Sr. Joarez Wernke, por ter cedido o conjunto de dados utilizado na aplicação 3.

Aos membros da banca em especial aos professores Oscar Dalfovo, Wilton de Oliveira Bussab e Mauto Rosemberg pelas contribuições e orientações que enriqueceram do trabalho.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

AAS	Amostragem Aleatória Simples
AE	Amostragem Estratificada
AEpr	Amostragem Estratificada Proporcional
AEun	Amostragem Estratificada Uniforme
ANSI	American National Standards Institute
EQM	Erro Quadrático Médio
ISO	International Standards Organization
KDD	Descoberta de conhecimento em base de dados (Knowledge Discovery in Database)
MD	Mineração de Dados (Data Mining)
MDL	Data Manipulation Language
SGBD	Sistemas Gerenciador de Banco de Dados
SIG	Sistema de Informações Gerenciais
SQL	Linguagem de consulta estruturada (Structured Query Language)
SPT	Sistemas de Processamento Transações

RESUMO

Este trabalho apresenta um algoritmo para identificar características que possam ser utilizadas num processo de amostragem estratificada. O algoritmo localiza as características e os seus respectivos valores que dividem o conjunto de dados em estratos, de tal forma que a variância do estimador, de uma média ou proporção, seja inferior à variância do estimador baseado em uma amostra aleatória simples. O algoritmo implementa o cálculo da variância do estimador baseado nos três métodos de alocação: uniforme, proporcional e alocação ótima de Neyman com custo fixo. Foi também implementado um novo método denominado GRD, baseado no princípio do ganho de informação, que exige menos recursos de processamento. O algoritmo foi aplicado em um conjunto de dados simulados para produzir estratos pré-definidos, e também, em um conjunto de dados real. Além disso, o algoritmo foi implementado parcialmente em um Gerenciador de Banco de Dados.

ABSTRACT

This work presents an algorithm developed to identify characteristics that can be used to define strata in a stratified sampling process. The algorithm finds the characteristics, and its respective values, that split the data set into strata, in such a way that the variance of the estimator, of the mean or proportion, is smaller than the variance of the estimator based on a simple random sampling process. The algorithm implements the calculation of the variance of the estimator based on the three methods of allocation: uniform, proportional and Neyman optimum allocation with fixed cost. It has also implemented a new method called GRD, based on the principle of the information gain, that demands less amount of computational processing. The algorithm was applied in a simulated data, built to produce well defined strata, and in a real data set. Moreover, the algorithm was partially implemented in a Data Base Management System (DBMS).

1 INTRODUÇÃO

1.1 Considerações Iniciais

Os sistemas de Processamento de Transações (SPT) registram não apenas as transações comerciais tradicionais, mas qualquer evento passível de ser controlado e monitorado (STAIR, 1996). Por outro lado, novas tecnologias têm criado dispositivos de armazenamento com capacidade cada vez maior, permitindo que todas estas informações sejam armazenadas em grandes bases de dados.

Uma área interdisciplinar específica, Knowledge Discovery in Databases (KDD) – Descoberta de Conhecimento em Banco de Dados, surgiu em resposta à necessidade de novas abordagens e soluções para viabilizar a análise destas grandes bases de dados. Particularmente, KDD tem obtido sucesso na área de marketing, onde a análise de banco de dados de clientes revela padrões de comportamento e preferências que facilitam a definição de estratégias de vendas. (FAYYAD, 1996). Uma etapa importante do KDD é a Mineração de Dados (MD), onde são aplicados algoritmos para a descoberta de conhecimento.

"Não há um método de Mineração de Dados que seja “universal” e a escolha de um algoritmo para uma aplicação particular é de certa forma uma arte". (FAYYAD, 1996) Para encontrar respostas ou extrair conhecimento, existem diversas técnicas de MD disponíveis na literatura (CHEN, 1996). As principais podem ser agrupadas em:

- Análise de agrupamentos;
- Análise de regressão;
- Análise discriminante;
- Árvores de decisão;
- Indução e/ou extração de regras;
- Redes neurais;
- Algoritmos evolucionários;
- Conjuntos difusos.

Para a escolha da técnica mais adequada, é estratégico se ter conhecimento prévio à respeito do problema do domínio da aplicação de MD, como por exemplo, quais são os atributos importantes, quais são os relacionamentos possíveis, o que é uma função útil para o usuário, que padrões já são conhecidos e assim por diante. Na

definição de Fayyad, a busca de conhecimento é descrita como um processo composto por várias etapas, incluindo: preparação dos dados, busca de padrões, avaliação do conhecimento e refinamentos.

Conforme Han e Kamber (2000), a amostragem pode ser utilizada no pré-processamento para reduzir o tamanho do conjunto de dados, pois a grande quantidade de dados pode inviabilizar a realização do processo de KDD, considerando que alguns algoritmos de MD processam apenas um número limitado de registros. Assim, utilizam-se técnicas de amostragem de dados reduz-se o tamanho do conjunto de dados, obtendo-se um subconjunto relevante e representativo para todo o conjunto de dados.

Os dados que fornecem a base para a realização do processo de KDD podem ser de diversas origens e são classificadas em origens internas ou externas. Os dados de origem interna normalmente são fornecidos por repositórios de dados alimentados pelos sistemas transacionais e, geralmente, são constituídas por Data Warehouse. Os dados de origem externa são obtidos fora do domínio da aplicação.

O processo de amostragem já faz parte do cotidiano. Quando um cozinheiro experimenta a sopa ou quando alguém faz exame de sangue, coleta-se uma amostra para fazer inferência sobre o todo. Porém não se pode ignorar o fato de que o cozinheiro tem que mexer bem o molho, tornando assim o molho mais homogêneo antes de provar (BOLFARINE e BUSSAB, 2005).

A não utilização de técnicas de amostragem adequadas pode gerar um subconjunto de dados de características não representativas, comprometendo as análises que não representariam a verdadeira situação dos fatos registrados. Uma das técnicas de amostragem que tem fornecido bons resultados é a amostragem estratificada, a qual trabalha com a divisão da população em subdivisões ou estratos. A estratificação será mais eficiente quanto mais homogêneos forem os estratos, em relação a uma determinada característica de interesse.

A obtenção de uma amostra representativa pode também ajudar as aplicações utilizadas junto a banco de dados relacional. O modelo relacional é o modelo mais aplicado no momento, sendo utilizado por grandes empresas e instituições para gerenciar e manter seus dados (ABBEY, 2002; RAMALHO, 1999; SCHERER, 2000). Nestas aplicações a extração de informações é realizada pela execução de consultas (*queries*).

A obtenção de informações gerenciais frequentemente utiliza gráficos, por se tratar de um instrumento de fácil interpretação. A elaboração de gráficos não requer que os dados sejam absolutamente exatos. É possível simplificar o processo de obtenção de informações gerencias aplicando técnicas de amostragem. (GARRISON, 2001), na qual infere-se sobre o todo sem ser necessário lidar com todo o conjunto de dados.. Em banco de dados, utiliza-se o recurso de visões (*views*) em substituição a todo o conjunto de dados. Portanto, o procedimento de extração de dados gerenciais pode ser realizado a partir de *views* de banco de dados, que fazem uso de amostras.

Neste trabalho, propõe-se um algoritmo para identificar, a partir de conjunto de dados, as características mais apropriadas para criação de estratos homogêneos em relação a uma característica de interesse.

1.2 Objetivos da Pesquisa

1.2.1 Objetivo Geral

A proposta geral deste trabalho é pesquisar e propor um método de identificação de estratos, baseado em uma pesquisa piloto ou base anterior e/ou utilizando alguma *proxy*[1] como característica de interesse.

1.2.2 Objetivos Específicos

Para se chegar ao objetivo geral é necessária a execução dos seguintes objetivos específicos:

- estudar os métodos de amostragem;
- desenvolver algoritmos que identifiquem características homogêneas.
- projetar e implementar um protótipo;
- comparar métodos e processos para categorização de dados contínuos;
- estudar critérios de parada e/ou poda;
- aplicar o algoritmo em um conjunto de dados real;

[1] Proxy é utilizada neste caso para definir uma característica equivalente ou representante da característica de interesse..

1.3 Organização do Trabalho

Este trabalho está organizado em seis capítulos. O primeiro capítulo apresenta o contexto em que o trabalho atua, definindo seus objetivos gerais e específicos, metodologia e trabalhos correlatos.

O segundo capítulo apresenta a fundamentação básica de amostragem, enquanto o terceiro capítulo apresenta o algoritmo proposto. Sua demonstração e aplicabilidade são apresentadas no quarto capítulo. No quinto capítulo são feitas considerações finais e propostas de trabalhos futuros.

2 TÓPICOS DE AMOSTRAGEM

2.1 Introdução

O objetivo do processo amostral, segundo Bolfarine e Bussab (2005) é: ”obter informação sobre o todo, baseado no resultado de uma mostra”.

A observação de todos os elementos ou indivíduos da população (censo) é, na maioria das situações, impossível de efetuar, quer por questões de tempo e custos, quer por questões operacionais de implementação. Para fazer face à crescente necessidade de informação, tanto por parte das empresas e instituições, surgiu a necessidade de desenvolver métodos estatísticos que permitissem recolher essa informação a partir da observação de apenas uma parte da população. De um modo geral, o termo amostragem é utilizado para designar um conjunto de técnicas estatísticas que permitem inferir sobre determinadas características ou parâmetros da população ou universo, a partir de um conjunto limitado dos seus elementos (amostra).

Erros que derivam de uma amostragem são essencialmente de dois tipos: os erros devidos à amostragem e os erros que não se devem à amostragem. Os erros que não estão relacionados com o processo de amostragem designam-se erros não amostrais e podem ocorrer em qualquer fase da implementação da amostragem. Alguns exemplos deste tipo de erros são: os erros da base de amostragem (problemas de cobertura, informação auxiliar incorreta ou desatualizada, ...); erros para obtenção da informação (defeitos do questionário, erros no registro das respostas, não resposta total ou parcial, ...); erros no processamento dos dados (edição, codificação, análise, ...).

A qualidade dos resultados de uma amostragem depende, assim, da qualidade com que todas as suas etapas são implementadas. Ao longo da dissertação, serão utilizados os termos planos de amostragem ou desenho da amostra para referir genericamente a forma como a amostra foi selecionada.

Existem duas categorias de seleção dos elementos de uma amostra: os métodos probabilísticos e os métodos empíricos (SÄRNDAL et al., 1992) Nas sessões que se seguem, faz-se uma descrição resumida destes métodos e das principais etapas para implementação de uma amostragem probabilística. Este capítulo tem por objetivo apresentar o enquadramento teórico necessário para a compreensão da metodologia

apresentada em capítulos posteriores. Os métodos de amostragem empíricos encontram-se fora do âmbito da dissertação e só serão comentados brevemente.

Na Seção 2.3 introduz-se a notação e as definições essenciais da teoria da amostragem. Em seguida apresentam-se os planos de amostragem: amostragem aleatória simples e amostragem aleatória estratificada.

2.2 Métodos de amostragem empíricos

Conforme descrito por Bolfarine e Bussab (2005) métodos de amostragem empíricos ou não probabilísticos são especialmente utilizados em amostragem de opinião e estudos de mercado e caracterizam-se pelo fato de não ser possível a priori determinar a probabilidade de um elemento pertencer à amostra. A facilidade da implementação destes métodos e a flexibilidade de seleção dos elementos da amostra permitem reduzir os custos e efetuar mais rapidamente a amostragem. No entanto, os métodos empíricos têm a grande desvantagem de não ser possível avaliar a qualidade dos resultados.

2.3 Métodos de amostragem probabilísticos

O princípio de base dos métodos probabilísticos é que a probabilidade de se selecionar um elemento da população para a amostra é conhecida. SÄRNDAL et al (1992) apresenta quatro condições necessárias para a obtenção de uma amostra probabilística de uma determinada população:

- Ser possível definir o conjunto de todas as amostras, $S = \{s_1 , s_2 , ..., s_m \}$, que se podem obter através do procedimento de amostragem;
- Ser conhecida a probabilidade $p(s)$ de selecionar a amostra s, do conjunto de amostras possíveis;
- Não ser nula a probabilidade de selecionar cada elemento da população;
- O processo de seleção dos elementos da amostra deve ser aleatório, não pode ser baseado em julgamentos empíricos, tal que cada amostra s que se pode obter tenha exatamente a probabilidade $p(s)$.

Neste contexto, designa-se formalmente por plano de amostragem a função $p(.)$ que define a distribuição de probabilidade sobre o conjunto $S = \{s_1 , s_2, ..., s_m \}$. O plano de amostragem irá determinar as propriedades estatísticas dos estimadores (por

exemplo, o valor esperado e a variância) que permitem avaliar a qualidade das estimativas obtidas.

Somente no caso das amostragens probabilísticas é possível apresentar medidas de exatidão ou precisão das estimativas obtidas a partir da amostra.

2.4 Planejamento e implementação de uma amostragem

A concepção de uma amostragem é um processo que envolve diversas fases interdependentes, onde estão claramente definidos os conceitos, métodos e procedimentos. Särndal, Swensson e Wretman (SÄRNDAL et al., 1992) apresentam as principais etapas para a concepção e implementação de uma amostragem:

- especificação do objetivo (pergunta da pesquisa) da amostragem;
- tradução do problema em estudo num problema de amostragem.;
- especificação da população alvo, características de interesse, variáveis auxiliares disponíveis e parâmetros a estimar;
- construção ou obtenção da base de amostragem;
- inventariar os recursos disponíveis em termos orçamentais, humanos, técnicos, de equipamentos, entre outros;
- especificação de requisitos a que a amostragem deve obedecer, como por exemplo, o cronograma e a precisão das estimativas;
- especificação do método de coleta dos dados, incluindo a elaboração do questionário;
- especificação do desenho da amostra (plano de amostragem), mecanismo de seleção da amostra e determinação da sua dimensão;
- especificação dos métodos de processamento dos dados, incluindo a edição e imputação;
- especificação da forma dos estimadores e das medidas de precisão;
- treino dos recursos humanos e organização do trabalho de campo;
- alocação de recursos às diferentes operações da amostragem;
- alocação de recursos ao controle e avaliação.

Existem diversas formas de amostragem, porém no escopo deste trabalho somente serão consideradas as amostragens aleatória simples e estratificada. Para

maiores informações sobre as diversas formas de amostragem, consultar Särndal et al. (1992) e Bolfarine e Bussab (2005), por exemplo.

2.5 Amostragem Aleatória Simples (AAS)

Amostragem aleatória simples (AAS) é o método mais simples e mais importante para seleção da amostra. É um processo de selecionar uma amostra de *n* elementos, dentre *N* elementos, de modo que cada uma das possíveis amostras dos elementos tenha a mesma probabilidade de ser selecionada (COCHRAN, 1977).

2.5.1 Notação e Definições

Seja uma população de *N* elementos assim representada: $U = \{1,2,...N\}$, universo de interesse, *Y* uma característica de interesse medida em cada unidade,

$$\overline{Y} = \mu = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad (2.1)$$

a média populacional e

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \mu)^2 \qquad (2.2)$$

a variância populacional.

Nos casos em que o interesse é na presença ou não de determinado atributo em cada unidade da população, define-se $Y_i = 1$ caso a unidade possua o atributo e $Y_i = 0$, caso contrario. Nestes casos, os parâmetros de interesse são:

$$P = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad (2.3)$$

a proporção populacional de elementos com a mesma característica,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - P)^2 = P(1-P) \qquad (2.4)$$

a variância populacional.

2.5.2 Estimadores e suas variâncias

A partir de uma amostra aleatória simples *S*, de tamanho *n*, utiliza-se a média amostral:

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i \in s} Y_i \qquad (2.5)$$

como o estimador da média populacional μ, cuja variância é dada por:

$$Var(\bar{y}) = \frac{\sigma^2}{n} \qquad (2.6)$$

Nos casos em que *Y* representa a presença de um determinado atributo, utiliza-se a proporção amostral definida por:

$$\hat{P} = \frac{1}{n}\sum_{i \in S} Y_i \qquad (2.7)$$

como o estimador da proporção populacional *P*, cuja variância é dada por:

$$Var(\hat{P}) = \frac{P(1-P)}{n} \qquad (2.8)$$

As expressões para as variâncias dos estimadores são apropriadas para amostragem aleatória simples com reposição ou sem reposição com *N* grande.

Para exemplificar, considerar a população hipotética apresentada na Tabela 1 e na Figura 1, a qual contém *N*=8 elementos e a informação do salário, em salários mínimos, e da região na qual a pessoa reside.

Tabela 1 – Conjunto de dados hipotéticos demonstração.

Unidade	Salário	Região
1	5	A
2	13	A
3	6	A
4	6	A
5	10	D
6	12	D
7	17	D
8	19	D

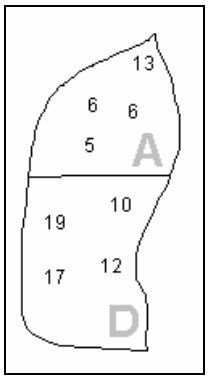

Figura 1 - Exemplo de conjunto de dados

A média da população é dada por (2.1):

$$\overline{Y} = (13+6+6+5+10+19+12+17)/8 = 11,$$

e a variância por (2.2):

$$\sigma^2 = [(13-11)^2+(6-11)^2+(6-11)^2+(5-11)^2+ (19-11)^2+(10-11)^2+(12-11)^2+(17-11)^2]/8 = 24.$$

Para uma amostra aleatória simples com reposição de tamanho n=4, a variância do estimador da média $\overline{y}$ é dada por (2.6):

$$Var(\overline{y}) = 24/4 = 6.$$

2.6 Amostragem Estratificada (AE)

A estratificação consiste na divisão da população em grupos ou estratos segundo uma ou mais características conhecidas dos elementos da população. A amostragem estratificada (AE) é a seleção probabilística de elementos de cada um dos estratos.

As principais razões para se utilizar a estratificação são (COCHRAN, 1977):

- Quando se desejam dados de uma determinada precisão sobre certas subpopulações.
- Conveniências administrativas, quando se deseja obter informação sobre uma determinada característica, como, por exemplo, desejamos obter dados para um estado ou uma região.
- Quando a estratificação proporcionar um aumento da precisão nas estimativas das características. Talvez seja possível dividir uma população heterogênea em subpopulações de tal forma que sejam mais homogêneas.

2.6.1 Notação e Definições

Seja um universo de *N* elementos *U*, definido na Seção 2.5, particionado em *H* partições, estratos, $U_1,...,U_h,\ ....,U_H$, isto é,

$$U = \bigcup_{h=1}^{H} U_h \text{ e } U_h \bigcap U_{h'} = \phi, \text{ com } h \neq h'.$$

Para que seja possível indicar convenientemente os elementos da população nos diferentes estratos, reescreve-se o universo a partir de:

$U_h = \{(h,1),....,(h,i),...,(h,N_h)\}$.

De modo análogo, características populacionais serão identificadas por dois índices: $Y_{h1},...,Y_{hN_H}$.

Os parâmetros populacionais média, proporção e variância são redefinidos, respectivamente, como:

$$\overline{Y} = \mu = \sum_{h=1}^{H} W_h \mu_h \qquad (2.9)$$

onde $W_h = \frac{N_h}{N}$: peso (proporção) do estrato *h*, com $\sum_{h=1}^{H} W_h = 1$, e

$\mu_h = \overline{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$: média do estrato *h*,

$$P = \sum_{h=1}^{H} W_h P_h \qquad (2.10)$$

onde P_h é a proporção de elementos no estrato *h* com o atributo desejado, e

$$\sigma^2 = \sum_{h=1}^{H} W_h \sigma_h^2 + \sum_{h=1}^{H} W_h (\mu_h - \mu)^2 \qquad (2.11)$$

onde $\sigma_h = \frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - \mu_h)^2$ é a variância do estrato *h.* Pode-se notar que a variância populacional depende não só das variâncias dos estratos como também das diferenças entre as médias dos estratos.

2.6.2 Efeito do Plano Amostral

Em 1965 Kish desenvolveu um método para comparar ganhos ou perda de precisão sob diferentes planos amostrais no estágio de planejamento da pesquisa. Esta medida é conhecida como Efeito do Plano Amostral (EPA ou DEFF –Design Effect) (PESSOA, 1998).

O EPA de Kish equivale a razão entre a variância do estimador obtido de um plano qualquer com relação a um plano estabelecido como padrão, geralmente o plano amostragem aleatório simples, cuja expressão é dada por:

$$EPA = \frac{Var_{aval}[\overline{y}]}{Var_{AAS}[\overline{y}]} \qquad (2.13)$$

2.6.3 Estimadores e suas variâncias

A partir de uma amostra aleatória simples, de tamanho n_h, de cada um dos estratos, utiliza-se a média ponderada:

$$\overline{y}_{es} = \sum_{h=1}^{H} W_h \overline{y}_h \qquad (2.12)$$

onde $\overline{y}_h$ é a média amostral do estrato *h*, como o estimador da média populacional μ, cuja variância é dada por:

$$Var(\overline{y}_{es}) = \sum_{h=1}^{H} W_h^2 Var(\overline{y}_h) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h} \qquad (2.13)$$

Nos casos em que *Y* representa a presença de um determinado atributo, utiliza-se a proporção amostral ponderada, definida por:

$$\hat{P}_{es} = \sum_{h=1}^{H} W_h \hat{P}_h \qquad (2.14)$$

onde $\hat{P}_h$ é a proporção amostral no estrato *h*, cuja variância é dada por:

$$Var(\hat{P}_{es}) = \sum_{h=1}^{H} W_h^2 Var(\hat{P}_h) = \sum_{h=1}^{H} W_h^2 \frac{P_h(1-P_h)}{n_h} \qquad (2.15)$$

Para exemplificar a utilização da estratificação, pode-se considerar que a população dada na Tabela 1 e Figura 1 tenha sido particionada em dois estratos

definidos pelas regiões A e D. Na Tabela 2 tem-se os parâmetros dos estratos e também os parâmetros populacionais.

Tabela 2- Exemplo de estratificação.

Estrato	Região	N_h	μ_h	σ_h^2
1	A	4	7,50	10,25
2	D	4	14,75	11,19
Total Populacional		8	11	24

Pode-se observar que com esta estratificação, foram obtidos estratos com médias diferentes e variâncias próximas e bem menores do que a variância populacional. No caso de uma amostra estratificada de *n=4* elementos, sendo 2 de cada estrato ($n_1 = n_2 = 2$), tem-se $W_1 = W_2 = 4/8 = 1/2$ e a variância do estimador da média dada por (2.13):

$$Var(\bar{y}_{es}) = (1/2)^2 \times \frac{10,25}{4} + (1/2)^2 \times \frac{11,19}{4} = 10,72 \ .$$

Comparado com a variância do estimador da média populacional obtido com uma AAS, este valor é bem menor indicando que este processo de estratificação fornece resultados mais precisos. Isto não quer dizer que sempre a AE fornecerá melhores resultados, para a estimação da média e da proporção populacional, do que a AAS. Para um mesmo tamanho de amostra, a AE será mais eficaz do que a AAS quanto maior for a habilidade do pesquisador em produzir estratos homogêneos (BOLFARINE e BUSSAB, 2005). Como pôde ser visto no exemplo discutido, as variâncias dos estratos, 10,25 e 11,19, são bem menores do que a variância populacional, 24, isto é, os estratos são bem mais homogêneos do que a população.

No exemplo apresentado, o número de elementos amostrados em cada estrato foi o mesmo. Dependendo do interesse do pesquisador e das características dos estratos, diferentes tamanhos de amostra podem ser utilizados nos estratos. A seguir são apresentados três métodos de alocação de uma amostra de tamanho *n* pelos estratos.

2.7 Alocação da Amostra pelos Estratos

Alocação da amostra é a distribuição da quantidade *n* de elementos da amostra nos estratos. Esta distribuição é muito importante, pois ela é que irá determinar a precisão do procedimento amostral (BOLFARINE e BUSSAB, 2005).

2.7.1 Alocação Proporcional

Nesta alocação, a quantidade de elementos amostrados *n* é dividida proporcionalmente ao tamanho do estrato, ou seja,

$$n_h = nW_h = n\frac{N_h}{N} \qquad (2.16)$$

Utilizando este valor em (2.15), pode-se mostrar que a variância do estimador da média populacional, obtido com a Amostragem Estratificada Proporcional (AEpr), é dada por:

$$Var(\bar{y}_{es}) = (\frac{1}{N}\sum_{n=1}^{H} N_h\sigma_h^2)/n \qquad (2.17)$$

2.7.2 Alocação Uniforme

Na alocação uniforme, a quantidade de elementos amostrados *n* é igualmente distribuída nos estratos, ou seja,

$$n_h = \frac{n}{H} = k \qquad (2.18)$$

Utilizando este valor em (2.15), pode-se mostrar que a variância do estimador da média populacional, obtido com a Amostragem Estratificada Uniforme (AEun), é dada por:

$$Var(\bar{y}_{est}) = (\frac{H}{N^2}\sum_{n=1}^{H} N_h^2\sigma_h^2)/n \qquad (2.19)$$

Este método de alocação é muito utilizado quando se pretende apresentar estimativas separadas para cada estrato.

2.7.3 Alocação Ótima de Neyman

Atribuída a Neyman (NEYMAN, 1934), ela consiste em determinar os n_h de modo que a variância do estimador seja mínima, para um determinado custo total fixo. Além do número e do tamanho dos estratos, os valores de n_h irão depender também do custo por unidade amostrada e da variância do estrato *h.* Supondo o mesmo custo por unidade observada em todos estratos, o número de elementos que deverá ser amostrado no estrato *h* é dado por:

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h} \qquad (2.20)$$

Utilizando este valor em (2.15), pode-se observar que a variância do estimador da média populacional, obtido com a Amostragem Estratificada com Alocação Ótima de Neyman, com custo fixo, é dada por:

$$Var(\bar{y}_{es}) = \left[\left(\frac{1}{N} \sum_{n=1}^{H} N_h \sigma_h \right) \right]^2 / n \qquad (2.21)$$

Para mais detalhes sobre critérios de estratificação e sobre estes métodos de repartição da amostra, em particular, ver (BARNETT, 1974; HEDAYAT, 1991).

3 ALGORITMO

O algoritmo foi desenvolvido com o objetivo de encontrar, em um conjunto de dados, as características, e suas respectivas classes, que possam melhor estratificar o universo de interesse, ou seja, encontrar características e classes que definam estratos homogêneos. A homogeneidade dos estratos é definida a partir de suas variâncias associadas a uma característica ou variável de interesse.

O procedimento é ilustrado no fluxograma da Figura 2. Pode-se observar que, para cada característica candidata, buscam-se suas classes distintas e calculam-se a variância conforme a forma de alocação definida previamente (ver Seção 3.1). A característica e suas classes que retornarem a menor variância é selecionada para a estratificação. O algoritmo também pode utilizar o método GRD (ver Seção 3.2) como alternativa à utilização da variância do estimador.

O fluxograma pode ser acompanhado pela Figura 2, onde se pode observar que o primeiro passo é obter uma amostra ou a base com dados que serão usados nos demais passos. No passo seguinte verificam-se as características candidatas, isso acontece, pois há possibilidade do pesquisador ter uma quantidade de características, mas queira selecionar apenas algumas destas. No passo seguinte, em posse da característica, buscam-se os valores distintos desta característica, e para cada uma é calculada a variância relacionada com a variável de interesse. O cálculo da variância depende do método de alocação selecionado ou do GRD. Os próximos passos são para verificar se o valor calculado é o menor. Caso isso aconteça então são armazenados a característica e o valor. Depois de realizar a procura de todas as características e valores o resultado é a característica com a menor variância.

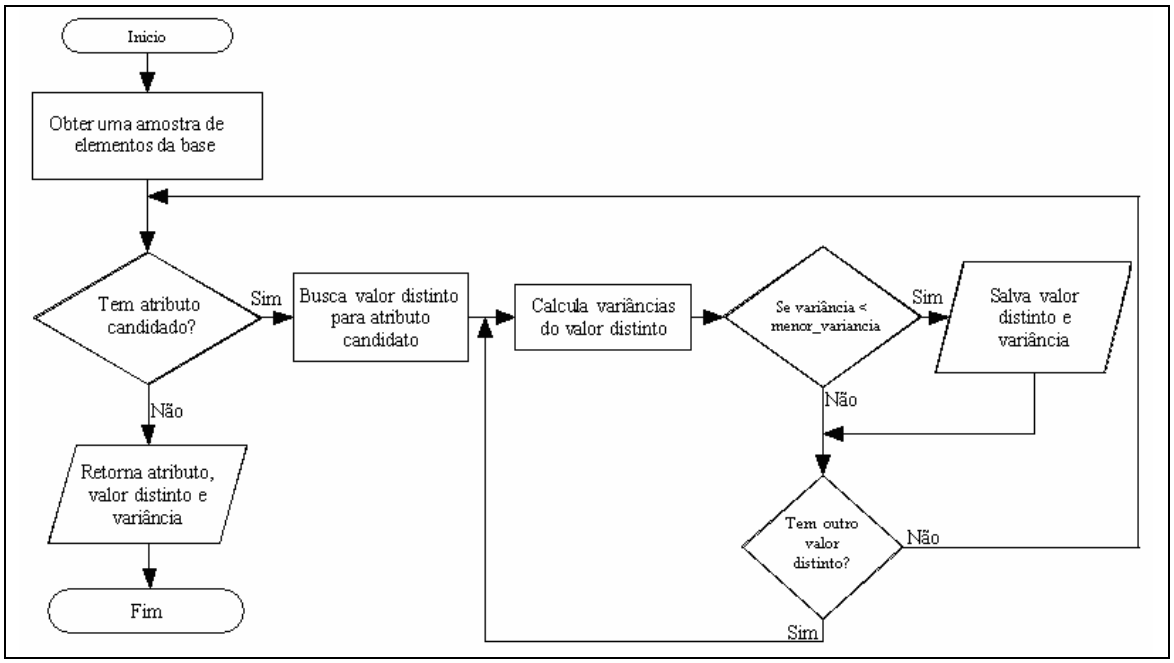

Figura 2 - Fluxograma da execução do algoritmo.

Este procedimento é executado de maneira recursiva, como pode ser visto no Quadro 1, ou seja, ao ser encontrada uma característica, classe para a estratificação, o conjunto de dados é dividido em duas partes. Em seguida, o algoritmo é aplicado em cada uma das partes, conforme pode ser visto no Quadro 1, até que o critério de parada seja atingido.

```
Função IdentificaEstratos (A atributos, E exemplos);
Inicio
     Se E está vazio, retorna falha
     Se E contém casos com o mesmo valor da variável de interesse
        Conclui Estrato

     Para cada variável (exceto a de interesse) em A faça
         Calcula valor da Variância (conforme método de alocação) de A[i]
     fim para

     A[x] = Atributo com menor variância
     Adicionar característica da variável A[x] ao estrato

     Para cada valor da variável
         chamar Função IdentificaEstratos (A, Ex)
     Fim para
Fim.
```

Quadro 1- Algoritmo de pesquisa características/classes que tenham menor variância.

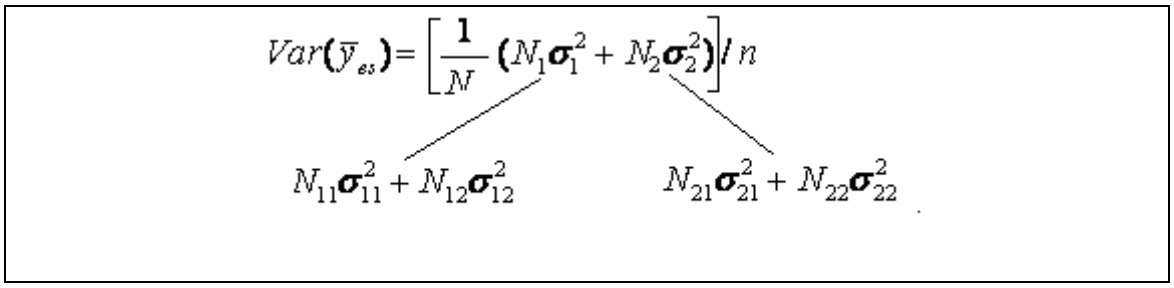

Figura 3 - Recursividade no cálculo da variância, com alocação proporcional.

Características contínuas são categorizadas em um número de categorias (faixas), *q*, que, necessariamente, deverá ser definido pelo usuário. As *q* faixas são construídas a partir do menor valor da característica *Min* e com amplitudes iguais, e dadas por: *(Max – Min) / q*, onde *Max* é a o maior valor da característica.

Foram implementados dois critérios de parada para finalizar a recursividade. Um dos critérios limita a quantidade máxima de características definidoras dos estratos e é válido para os três métodos (uniforme, proporcional e ótima). O outro critério define o ganho mínimo necessário para ser adicionada uma nova partição no processo de estratificação, isto é o percentual mínimo de redução na variância do estimador. Este critério somente é aplicável para o método que utiliza alocação.

3.1 Métodos Baseados na Alocação

A variância do estimador é calculada conforme o tipo de alocação definido previamente:

- Proporcional: expressão (2.17);
- Uniforme: expressão (2.19);
- Ótima: expressão (2.21).

Os resultados das variâncias são apresentados como função de *n*, de modo a proporcionar uma comparação direta com a variância do estimador obtido a partir de uma amostra aleatória simples, σ^2/n.

3.2 Método GRD

Nesta secção apresenta-se o método GRD que leva em conta o ganho de informação contido nas características pesquisadas e não a variância do estimador para identificação das características de estratificação.

3.2.1 Entropia e Ganho de Informação

Entropia é uma medida que indica o grau de heterogeneidade da classe de uma característica $A_{l,s}$ onde $l=1,2,...,L$ representa a característica l estudada e $s = 1,2,\ldots,a_{,}$, de uma característica A_l . Suas definições são expressas em bits e dadas por (SHANNON,1948):

$$Info(A_{l,s}) = -p(A_{l,s})\log_2 p(A_{l,s}) \qquad (3.1)$$

e

$$Info\,(A_l) = \sum_{s=1}^{a_l} Info\,(A_{l,s}) \qquad (3.2)$$

onde $p(A_{l,s})$ é a proporção de elementos com a classe s. A entropia tende a ficar próxima de zero quando os dados são mais homogêneos. Por exemplo, para a característica gênero, uma classe seria masculino e $p(A_{l,s})$ seria a proporção de elementos do sexo masculino.

O ganho de informação para uma classe s da característica l é dado por:

$$Gain(A_{l,s}) = Info(A_l) - Info(A_{l,s}) \qquad (3.3)$$

A expressão (3.3) pode produzir grandes distorções em favor das classes que ocorrem para quase todos elementos. Por isso, utiliza-se a expressão do ganho relativo, o qual é baseado no número de elementos obtido nas proporções, e dado por:

$$Gainratio(A_{l,s}) = \frac{Gain(A_{l,s})}{Split(A_{l,s})} \qquad (3.4)$$

onde

$$Split(A_{l,s}) = -\frac{T_{l,s}}{T_l} \times \log_2 \frac{T_{l,s}}{T_l} \qquad (3.5)$$

com $T_{l,s}$ representando o número total de elementos com a classe s para a característica l e T_l o número total de elementos com a característica l.

3.2.2 GRD

Neste trabalho propõe-se o método GRD, o qual se baseia em (3.1) e (3.5), e utiliza apenas a entropia com a proporção de casos de uma determinada classe da característica, enfatizando assim o ganho proporcionado por determinada classe, e definido como:

$$GRD(A_{l,s}) = \log_2 p(A_{l,s}) \times \log_2 \frac{T_{l,s}}{T_l} \qquad (3.6)$$

O processo de ganho de informação é calculado sobre a característica candidata e a característica de interesse e tem como objetivo determinar qual a característica que proporciona o maior ganho.

Para exemplificar o método GRD, considere os dados apresentados na Tabela 3, obtidos de Quilan (1986).

Tabela 3 - Base de "Jogo de Golfe" introduzido por Quilan.

Nr.	Tempo	Temperatura (ºF)	Umidade	Vento	Joga	
1	Sol	85	85	Não	Não	0
2	Sol	80	90	Sim	Não	0
3	Nublado	83	86	Não	Sim	1
4	Chuva	70	96	Não	Sim	1
5	Chuva	68	80	Não	Sim	1
6	Chuva	65	70	Sim	Não	0
7	Nublado	64	65	Sim	Sim	1
8	Sol	72	95	Não	Não	0
9	Sol	69	70	Não	Sim	1
10	Chuva	75	80	Não	Sim	1
11	Sol	75	70	Sim	Sim	1
12	Nublado	72	90	Sim	Sim	1
13	Nublado	81	75	Não	Sim	1
14	Chuva	71	91	Sim	Não	0

Para identificar a informação de ganho das características, como neste exemplo, para obter a informação de ganho da característica *Vento* sobre a variável de interesse que neste caso é a característica de interesse *Joga*.

Info(Joga) = [p(Joga, Sim) x log2(p(Joga, Sim)] + [p(Joga, Não) x
log2(p(Joga, Não)]

= [9/14 x log2(9/14)] + [5/14 x log2(5/14)]

Info(Joga) = 0,940285959

Vento:

D(vento=sim; joga=sim) = 3/6 x log2(3/6) = 0,5

D(vento=sim; joga=não) = 3/6 x log2(3/6) = 0,5

D(vento=não; joga=sim) = 6/8 x log2(6/8) = 0,311278124

D(vento=não; joga=não) = 2/8 x log2(2/8) = 0,811278124

Gain(Joga, Vento) = Info(Joga)–{[6/14 x (0,5+0,5)] + [8/14 x (0,3113 + 0,81128)]}

Gain (Joga, Vento) = 0,04813

Para as demais características tem-se:

Gain(Joga,Temperatura) = Info(Joga)–{[5/14x(0,2575+0,4644)]+[9/14 x (0,47111+ 0,52]}

Gain (Joga,Temperatura) = 0,045335

Umidade:

Gain (Joga, Umidade) = 0,102

Para a característica *Tempo*:

Gain (Joga, Tempo) = 0,2468

Conforme pode ser observado, a característica que possui a maior informação de ganho é a característica *Tempo*. Portanto, esta será a característica indicada com a sendo a característica de maior ganho de informação. Ao aplicarmos na mesma base a algoritmo o estrato sugerido pela forma de cálculo GRD será *(Vento = 'Não') E (Tempo = 'Chuva')*.

Aplicando-se o método GRD em todas as características do conjunto de dados, tem-se:

Vento:

GRD(vento=sim; joga=sim) = log2(6/9) x log2(6/14) = 1,937

GRD(vento=não; joga=sim) = log2(3/9) x log2(6/14) = 0,715

Tempo:

GRD(tempo=sol; joga=sim) = log2(2/9) x log2(5/14) = 3,223

GRD(tempo=chuva; joga=sim) = log2(3/9) x log2(5/14) = 2,354

GRD(tempo=nublado; joga=sim) = log2(4/9) x log2(4/14) = 2,114

Temperatura: característica continua na qual foram consideradas diferentes formas de categorização. Abaixo se apresenta aquela que forneceu o melhor resultado.

GRD(64 =< Temperatura =< 72); joga=sim) = log2(5/9) x log2(6/14)

= 1,036

Umidade: característica continua na qual foram consideradas diferentes formas de categorização. Abaixo se apresenta aquela que forneceu o melhor resultado.

GRD(65 =< umidade =< 75; joga=sim) = log2(4/9) x log2(5/14)

= 1,737

Observa-se que o menor valor foi obtido com a característica/classe *Vento=não*. Aplicando-se novamente o algoritmo, obtiveram-se as características/classes *(Vento = 'Não') e (Tempo = 'Chuva').* Na Tabela 4 são apresentados os resultados.

Tabela 4 – Resultados da aplicação do método GRD ao conjunto de dados Golfe.

Estratos (GRD)		Quantidade	Média	Variâncias
1	(Vento = 'Não') E (Tempo = 'Chuva')	3	1,00000	0,00000
2	(Vento = 'Não') E (Tempo <> 'Chuva')	5	0,60000	0,24000
3	(Vento <> 'Não') E (Tempo = 'Chuva')	2	0,00000	0,00000
4	(Vento <> 'Não') E (Tempo <> 'Chuva')	4	0,75000	0,18750
	Global	14	0.6429	0,2296

.

Na Tabela 5, tem-se os resultados das variâncias dos estimadores aplicando-se os diferentes tipos de alocação à estratificação proposta pelo método GRD.

Tabela 5 - Variâncias dos estimadores.

Alocação	Uniforme	Proporcional	Ótima	AAS
Variância	0,18367/n	0,13929/n	0,08921/n	0,2296/n

4 APLICAÇÕES

Este capítulo apresenta a aplicação do algoritmo para a identificação de características para estratificação em um conjunto de dados simulados e um conjunto de dados reais de clientes e vendas de um supermercado e uma outra aplicação demonstrando como aplicar o algoritmo diretamente em uma estrutura de banco de dados utilizando-se para a obtenção de consultas gerenciais.

O Protótipo do programa para aplicação do algoritmo foi construído utilizando a linguagem Delphi, e o conjunto de dados segue o formato Microsoft Access. Para sua execução é necessário informar o conjunto de dados, a característica de interesse, as características candidatas à estratificação, o critério de parada, o tipo de alocação e como deverão ser categorizadas as características contínuas. O Protótipo do programa, o programa fonte e a documentação podem ser acessados no sítio: www.inf.ufsc.br/~dandrade.

4.1 Conjunto de Dados Simulado

Para avaliar o funcionamento do algoritmo utilizou-se um conjunto de dados simulado, composta das características Sexo, Rendimento, Escolaridade e Altura, conforme apresentado na Tabela 4. O conjunto de dados foi elaborado de maneira que a característica *Rendimento* fosse bem distinta entre as classes Masculina e Feminina da característica *Sexo* e entre as classes Alta, Média e Baixa da característica *Escolaridade.* Com relação a característica *Altura*, seus valores foram gerados de modo a não serem correlacionados com a característica *Rendimento*. Foram gerados 6000 elementos com rendimento médio igual 51,66 e variância igual a 324,92. Detalhes nas tabelas 6 e 7.

Tabela 6 - Características do Conjunto de Dados Simulado.

Característica	Tipo	Classes
Sexo	Categórica	M = Masculina F = Feminina
Escolaridade	Categórica	A=Alta M= Média B=Baixa
Altura (em metros)	Contínua	1,41 a 2,14
Rendimento (em mil Reais)	Contínua	17 a 92

Tabela 7- Estatísticas dos dados simulados.

			Altura		Rendimento	
Estrato	Sexo	Escolaridade	Média	Variância	Média	Variância
1	Masculino	Baixa	1,75	0,0092	30,03	8,69
2	Masculino	Média	1,75	0,0102	29,90	8,72
3	Masculino	Alta	1,75	0,0106	49,41	10,52
4	Feminino	Baixa	1,75	0,0094	60,21	12,61
5	Feminino	Média	1,75	0,0089	60,06	13,08
6	Feminino	Alta	1,75	0,0105	79,84	13,49

A aplicação do algoritmo, independentemente do método utilizado, tipo de alocação ou GRD, produziu os mesmos estratos, conforme apresentado no Quadro 2. Entretanto, o método GRD foi em média, 24% mais rápido que os métodos de alocação.

```
Estrato: 1
    (SEXO = 'F') e (ESCOLAR = 'A')
   Variância  13,4882, Média 79,8414, Quantidade 1000

Estrato: 2
    (SEXO = 'F') e (ESCOLAR <> 'A')
   Variância  12,8461, Média  60,1365, Quantidade 2000

Estrato: 3
    (SEXO <> 'F') e (ESCOLAR = 'A')
   Variância  10,5714, Média 49,9119, Quantidade 1000

Estrato: 4
    (SEXO <> 'F') e (ESCOLAR <> 'A')
   Variância   8,7063, Média 29,9640, Quantidade 2000
```

Quadro 2 - Resultado do Conjunto de dados do simulado.

Pode-se observar a partir do Quadro 2 que há uma redução significativa da variabilidade na utilização da estratificação. Por exemplo, a alocação uniforme fornece uma variância igual a 12,25/n, enquanto que a variância do estimador utilizando AAS é igual a 324,92/n.

4.2 Conjunto de Dados Reais

Para demonstrar a aplicabilidade do algoritmo sob condições mais reais, foi utilizado um conjunto de dados de cliente e vendas de um supermercado. As características da base já preparadas para esta aplicação estão na Tabela 8. Alguns elementos da base podem ser vistos no Anexo 1.

Tabela 8 - Características dos clientes nas vendas do supermercado.

Característica	Descrição
QCProd1	Quantidade anual comprada do produto Abacaxi;
VVProd1	Valor anual pago pelo produto Abacaxi;
QCProd2	Quantidade anual comprada do produto Banana;
VVProd2	Valor anual pago pelo produto Banana;
QCProd3	Quantidade anual comprada do produto Laranja;
VVProd3	Valor anual pago pelo produto Laranja;
QCProd4	Quantidade anual comprada do produto Maçã;
VVProd4	Valor anual pago pelo produto Maçã;
QCProd5	Quantidade anual comprada do produto Mamão;
VVProd5	Valor anual pago pelo produto Maçã Mamão;
QCProd6	Quantidade anual comprada do produto Uva Itália;
VVProd6	Valor anual pago pelo produto Uva Itália;
QCProd7	Quantidade anual comprada do produto Uva Crismson;
VVProd7	Valor anual pago pelo produto Uva Crismson;
QCProd8	Quantidade anual comprada do produto Tomate Caqui;
VVProd8	Valor anual pago pelo produto Tomate Caqui;
QCProd9	Quantidade anual comprada do produto Banana;
VVProd9	Valor anual pago pelo produto Banana;
QCProd10	Quantidade anual comprada do produto Melão;
VVProd10	Valor anual pago pelo produto Melão;
RegiaoCliente	Indica a região do cliente;
Sexo	Indica o sexo do cliente;
GrauInstrucao	Característica categórica que determina o grau de instrução do cliente;
EstadoCivil	Estado civil do cliente.

O objetivo é verificar a aceitação de um novo produto, maça argentina, através de uma pesquisa junto a uma amostra de clientes cadastrados. O parâmetro de interesse a ser estimado é a proporção de clientes cadastrados que comprariam o novo produto, caso fosse oferecido pelo supermercado.

Para que não fosse necessário pesquisar todos seus clientes, ou ainda, para que não fosse utilizada uma amostra que não refletisse a realidade dos consumidores deste novo produto, e ainda considerando que este produto nunca foi comercializado pela empresa, foi utilizado uma *proxy* de um produto similar, a característica *QCProd4*, ou seja, quantidade anual de compras da maça nacional. Neste caso, deseja-se encontrar estratos formados por clientes homogêneos em relação a esta *proxy*. A média e a variância da característica *QCProd4* são, respectivamente, 0,00413 e 0,006566.

Assim, baseado no consumo de maçã brasileira, foi identificado quais os grupos (estratos) de pessoas que deveriam ser utilizados na seleção da amostra de clientes a ser pesquisada. Foram aplicados os diferentes métodos descritos no capítulo anterior, e os resultados obtidos estão apresentados no Anexo 2.

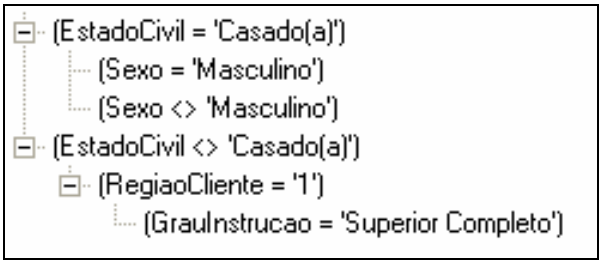

Figura 4 - Resultado da simulação.

A Figura 4 apresenta os resultado do algoritmo com a alocação uniforme. Pode-se observar que os estratos são apresentados na estrutura de uma árvore.

As árvores são amplamente usadas em processos de classificação, onde partindo da raiz da árvore para algum nó folha que fornece as características. Cada nó da árvore especifica uma característica, e cada arco que desce daquele nó corresponde a um dos possíveis valores destas características. Um estrato é obtido começando no nodo raiz da árvore e segue o arco que corresponde ao valor da característica. Este processo é repetido então para a sub-árvore abaixo até chegar a um nodo folha, ou no limite de ramos.

A interpretação obtida é que a primeira característica/classe identificada para a estratificação foi *EstadoCivil= 'Casado'*. Em seguida, o algoritmo sugere a estratificação pela característica *Sexo*. Para os clientes *não casados*, e pela característica *GrauInstrução*, para os clientes não casados pode ser visto na Figura 5. Em relação a esta última característica, os clientes devem ser estratificados em *Superior Completo* e outros.

Os estratos sugeridos que podem ser vistos na Figura 6, onde se se pode observa que a média de valores da característica *QCProd4* se destinge para classes de *EstadoCivil = Casado (a) e Sexo = Masculino*.

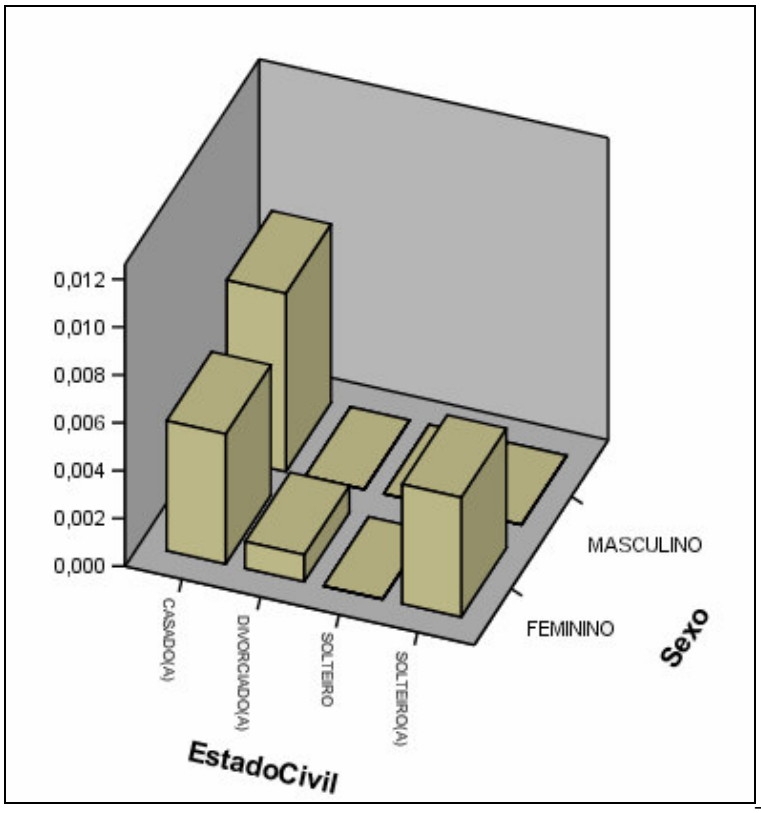

Figura 5 - Características obtidas na estratificação uniforme com Estado Civil igual Casado.

No outro ramo do estrato, que pode se ser visto na Figura 6, vê-se que para característica *EstadoCivil* diferente *Cadado (a)* a característica *RegiaoCliente* igual 1 difere das demais regiões.

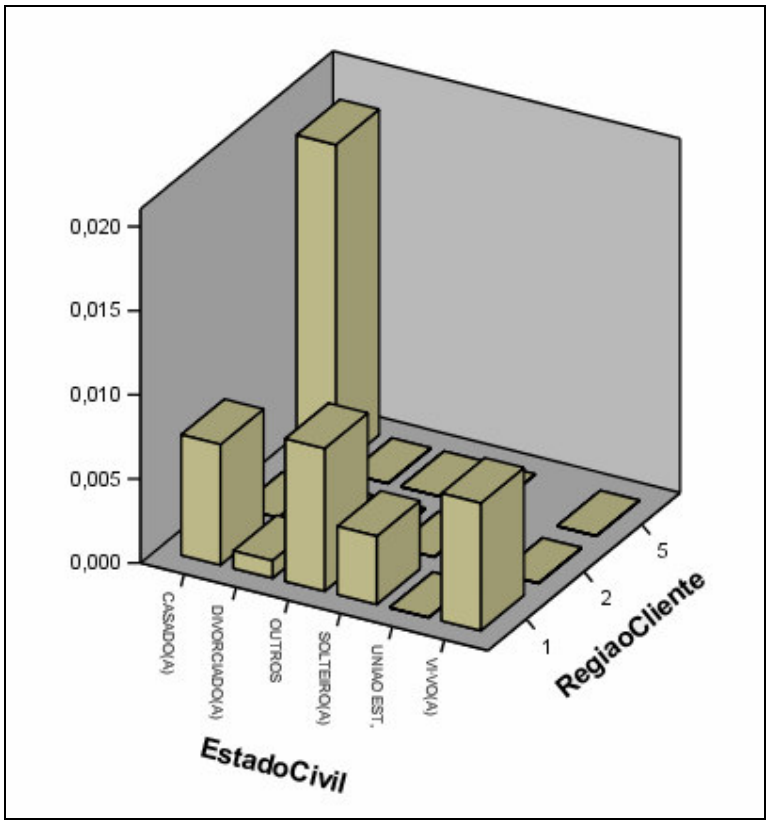

Figura 6 - Características obtidas na estratificação uniforme com EstadoCivil diferente Casado.

O resultado obtido dos estratos sugeridos pelo algoritmo, utilizando o critério de parada de limite máximo de quatro características por estrato, está demonstrado no Quadro 3.

```
RESUMO : Atributo meta :  média  0.0122, Casos  17165, variância  0.06566

Estrato: 1:  (EstadoCivil = 'Casado(a)') e (Sexo = 'Masculino') e (GrauInstrucao = 'Superior Completo') e (RegiaoCliente = '1')
  Variância  0.00654,  Média  0.00659, Quantidade 1062
Estrato: 2:  (EstadoCivil = 'Casado(a)') e (Sexo = 'Masculino') e (GrauInstrucao = 'Superior Completo') e (RegiaoCliente <> '1')
  Variância  0.02059, Média  0.01244, Quantidade 241
Estrato: 3:  (EstadoCivil = 'Casado(a)') e (Sexo = 'Masculino') e (GrauInstrucao <> 'Superior Completo') e (GrauInstrucao = 'Primeiro
Grau')       Variância  0.13547, Média  0.01244, Quantidade 124
Estrato: 4: (EstadoCivil = 'Casado(a)') e (Sexo = 'Masculino') e (GrauInstrucao <> 'Superior Completo') e (GrauInstrucao <> 'Primeiro
Grau')       Variância  0.00506, Média  0.00304, Quantidade 986
Estrato: 5:  (EstadoCivil = 'Casado(a)') e (Sexo <> 'Masculino') e (GrauInstrucao = 'Superior Completo') e (RegiaoCliente = '1')
  Variância  0.01516, Média  0.00944, Quantidade 1377
Estrato: 6: (EstadoCivil = 'Casado(a)') e (Sexo <> 'Masculino') e (GrauInstrucao = 'Superior Completo') e (RegiaoCliente <> '1')
  Variância  0.0000, Média  0.00000, Quantidade 297
Estrato: 7: (EstadoCivil = 'Casado(a)') e (Sexo <> 'Masculino') e (GrauInstrucao <> 'Superior Completo') e (GrauInstrucao = 'Segundo
Grau')        Variância  0.0030, Média  0.00308, Quantidade 648
Estrato: 8 : (EstadoCivil = 'Casado(a)') e (Sexo <> 'Masculino') e (GrauInstrucao <> 'Superior Completo') e (GrauInstrucao <> 'Segundo
Grau')       Variância  0.0029, Média  0.00300,  Quantidade 998
Estrato: 9: (EstadoCivil <> 'Casado(a)') e (RegiaoCliente = '1') e (GrauInstrucao = 'Superior Completo') e (EstadoCivil = 'Solteiro(a)')
  Variância  0.0155, Média  0.00726, Quantidade 964
Estrato: 10: (EstadoCivil <> 'Casado(a)') e (RegiaoCliente = '1') e (GrauInstrucao = 'Superior Completo') e (EstadoCivil <> 'Solteiro(a)')
Variância  0.00427, Média  0.00429,  Quantidade 931
Estrato: 11:  (EstadoCivil <> 'Casado(a)') e (RegiaoCliente = '1') e (GrauInstrucao <> 'Superior Completo') e (Sexo = 'Feminino')
  Variância  0.0039, Média  0.00394,  Quantidade 2028
Estrato: 12:  (EstadoCivil <> 'Casado(a)') e (RegiaoCliente = '1') e (GrauInstrucao <> 'Superior Completo') e (Sexo <> 'Feminino')
  Variância  0.0004, Média  0.00042,  Quantidade 2331
Estrato: 13: (EstadoCivil <> 'Casado(a)') e (RegiaoCliente <> '1') e (Sexo = 'Feminino') e (RegiaoCliente = 'REGIAO')
  Variância  0.0070, Média  0.00330,  Quantidade 2121
Estrato: 14: (EstadoCivil <> 'Casado(a)') e (RegiaoCliente <> '1') e (Sexo = 'Feminino') e (RegiaoCliente <> 'REGIAO')
  Variância  0.0000, Média  0.00000, Quantidade 678
Estrato: 15: (EstadoCivil <> 'Casado(a)') e (RegiaoCliente <> '1') e (Sexo <> 'Feminino') e (GrauInstrucao = 'GRAU INSTRUCAO')
  Variância  0.0042, Média  0.00431,  Quantidade 1390
Estrato: 16: (EstadoCivil <> 'Casado(a)') e (RegiaoCliente <> '1') e (Sexo <> 'Feminino') e (GrauInstrucao <> 'GRAU INSTRUCAO')
  Variância  0.0040, Média  0.00202, Quantidade 989
```

Quadro 3 - Estratos gerados pelo algoritmo.

Utilizando-se todos os estratos obtidos no Quadro 3, com alocação uniforme, obtive-se variância do estimador igual a 0,001735/n, bem inferior à aquela apresentada pela AAS 0,006566/n, com EPA igual a 0,26424, representando somente 26% da variância da AAS.

4.3 Relatórios Gerenciais

Nesse outro exemplo utilizou-se outro enfoque, aplicando a geração de estratos diretamente no SGBD (Sistema Gerenciador de Banco de Dados), objetivando obter visões de amostras de elementos a serem utilizadas na produção de consultas / relatórios gerenciais.

A maioria dos SGBDs possui uma Data Manipulation Language (DML) (COOD 1970) baseada no padrão SQL (Structured Query Language) definido pela ANSI (Americam National Standards Institute) e pela ISO (International Standards Organization) que é amplamente usada para geração de consultas e/ou relatórios sobre os dados mantidos pelos SGBDs. A linguagem SQL (é, atualmente, um padrão para

gerenciamento de dados em SGBDs relacionais. Os SGBDs de grande porte mais famosos são: Oracle, SQL Server, Informix, Sybase e Ingres (FANDERUFF, 2000, KORTH, 2001, YE 2003).

O modelo relacional é o modelo mais utilizado no momento. Grandes empresas e instituições utilizam SGBDs relacionais para gerenciar e manter seus dados (ABBEY, 2002; RAMALHO, 1999; SCHERER, 2000).

Um SIG (Sistema de Informações Gerenciais) baseia-se nos dados coletados pelos SPTs e normalmente envolve uma grande massa de dados. Por isso, geralmente as consultas e relatórios gerenciais consomem grande tempo para seu processamento. Por outro lado, não é exigido que as informações gerenciais sejam exatas, pois a maioria das informações é apresentada sob a forma de gráficos evidenciando alguma tendência (STAIR 1996).

Organizações com grande volume de dados utilizam um repositório (data warehouse) de informações coletadas de diversas fontes para elaboração de consultas gerenciais. Esta redundância visa evitar que seja afetado o desempenho dos SPTs durante o processamento das consultas gerenciais (SILBERSCHATZ, 1999).

A proposta desta seção é demonstrar como utilizar os recursos dos Gerenciadores de Banco de Dados para extração de consultas gerenciais baseando-se na técnica de amostragem estratificada, considerando os seguintes aspectos:

- Custo de processamento: ao trabalhar com uma amostra dos dados tem-se um custo de processamento menor;
- Precisão: com a utilização do método de amostragem estratificada, podem-se obter valores estimados próximos dos reais;
- Versatilidade: o método pode ser aplicado a uma série de diferentes aplicações e, portanto, torna-se independente de áreas de negócio;
- Complexidade: algoritmo simples que pode ser convertido em uma função ou *store procedure* da grande maioria dos gerenciadores de banco de dados (SGBD).
- Custo: custo reduzido pela pouca necessidade de poder de processamento dos equipamentos;

Para testes utilizou-se a extensão de linguagem SQL fornecida pelo MS SQL Server, versão 2000, ver Quadro 4 e Quadro 5. Nesta demonstração utilizou-se apenas alocação proporcional que pode ser vista na expressão (2.7).

```
create procedure Calcula_Variancia_Coluna(
@pColuna Varchar(30), -- Nome do atributo
@pTabela varchar(30), -- Nome da tabela
@pParam varchar(500), -- Filtro para concatenação
@PMeta Varchar(30)) -- variável de interesse
as
begin
(...)
   set @wMenorVar = 999999;
   set @wSql = N'set @wCursor = cursor local forward_only static read_only for
   select distinct(' +@pColuna+ ') from ' +@pTabela+ ' where (1=1) '+@pParam +
    ' ORDER BY ' + @pColuna+'; open @wCursor ';
   execute sp_executesql @wSql, N'@wCursor cursor output', @wCursor output
   if cursor_status('variable', '@wCursor') >= 0
   begin
      FETCH NEXT FROM @wCursor INTO @wValor;
      while (@@fetch_status=0)
      begin
        set @wVariancia = 0;
        execute db_teste.teste.Calcula_Variancia_campo_valor @pColuna,
           @pTabela, @pParam, @pMeta, @wValor, @wVariancia output;
      end
    end
(...)
end
```

Quadro 4 - Exemplo do Algoritmo aplicado sob o SGBD da MS – SQL.

```
create procedure Calcula_Variancia_Coluna(
@pColuna Varchar(30), -- Nome do atributo
@pTabela varchar(30), -- Nome da tabela
@pParam varchar(500), -- Filtro para concatenação
@PMeta Varchar(30)) -- variável de interesse
as
begin
(...)
   set @wMenorVar = 999999;
   set @wSql = N'set @wCursor = cursor local forward_only static read_only for
   select distinct(' +@pColuna+ ') from ' +@pTabela+ ' where (1=1) '+@pParam +
     ' ORDER BY ' + @pColuna+'; open @wCursor ';
   execute sp_executesql @wSql, N'@wCursor cursor output', @wCursor output
   if cursor_status('variable', '@wCursor') >= 0
   begin
      FETCH NEXT FROM @wCursor INTO @wValor;
      while (@@fetch_status=0)
      begin
         set @wVariancia = 0;
         execute db_teste.teste.Calcula_Variancia_campo_valor @pColuna,
                @pTabela, @pParam, @pMeta, @wValor, @wVariancia output;
         declare @wValorMaior float; --Armazena a valor da variância
         declare @wContadorMaior Numeric(8,0); --Nr. registros com esta
variância
         set @wSqlMenor = N'set @wCursorMenor = cursor local forward_only
static
         read_only for select var('+@pMeta+'), count('+@pMeta+') from '
              +@pTabela+ ' where (1=1) AND '+@pColuna+ ' <= '+Cast(@pvalor as
              Varchar)+' '+@pParam + ';
         open @wCursorMenor;';
         set @wSqlMaior = N'set @wCursorMaior = cursor local forward_only
static
         read_only for select var('+@pMeta+'), count('+@pMeta+') from '
                  +@pTabela+ '  where (1=1) AND '+@pColuna+ ' > '+Cast(@pvalor
as
                  Varchar)+' '+@pParam + ';
         open @wCursorMaior;';
         execute sp_executesql @wSqlMenor, N'@wCursorMenor cursor output',
         @wCursorMenor output
```

```
         if cursor_status('variable', '@wCursorMenor') >= 0
         begin
            FETCH NEXT FROM @wCursorMenor INTO @wValorMenor, @wContadorMenor;
         end;
         execute sp_executesql @wSqlMaior, N'@wCursorMaior cursor output',
             @wCursorMaior output
         if cursor_status('variable', '@wCursorMaior') >= 0
         begin
             FETCH NEXT FROM @wCursorMaior INTO @wValorMaior, @wContadorMaior;
         end;
         set @@wResultado =
            (power(@wContadorMenor/@wContadorMenor+@wContadorMaior,2)) *
                @wValorMenor +
                (power(@wContadorMaior/@wContadorMenor+@wContadorMaior,2)) *
                @wValorMaior;
             -- Calculo da Variância proporcional
         close @wCursorMenor;
         deallocate @wCursorMenor;
         close @wCursorMaior;
         deallocate @wCursorMaior;
      end
   end
end
```

Quadro 5 - Procedimento de cálculo da variância utilizando SQL da MS-SQL.

O conjunto de dados utilizado neste exemplo pode ser observado na Tabela 9, onde a característica de interesse é a *valorscompra* com a média de 23,56 e variância de 58,22.

Tabela 9 – Conjunto de dados utilizados em aplicação no BD.

Característica	Tipo	Classes
Sexo	Categórica	M = Masculina F = Feminina
Cidade	Categórica	Cidades de SC.
Idadecliente	Contínua	15 a 70
vlrvenda (valor de venda) (em Reais)	Contínua	0 a 100

Como resultados, criam-se comandos em linguagem SQL, (Quadro 6) como, por exemplo, uma *view*, para ser utilizada para obter uma coleção reduzida de elementos, mas mantendo uma acurácia sobre a variável de interesse.

```
create view viewvendas as
select top 900 vlrvenda from tabvendas where idadecliente <= 35
union all (
select top 100 vlrvenda from tabvendas where idadecliente > 35 )
```

Quadro 6 – Visão (*View)* criada pelo algoritmo.

Neste caso, a variável de interesse trata-se do atributo *vlrvenda* (valor de venda). Então se aplica o algoritmo para minimizar o acesso ao banco de dados. Com isso foi identificado que existem dois grupos homogêneos: um pertencente aos clientes com idade inferior a 35 anos e outro de clientes com idade igual ou superior a 35 anos. Com a utilização desta nova *view,* a variância resultante dá $23/n$, com EPA igual a 0,39505, representando somente 39% da variância da AAS.

5 CONCLUSÃO

Neste trabalho foi apresentado um algoritmo para identificação de estratos, cuja finalidade é auxiliar os pesquisadores na obtenção de planos amostrais estratificados mais condizentes com o conjunto de dados. Foi implementado um protótipo para utilização do mesmo e também foi demonstrada sua utilização diretamente no Gerenciador de banco de dados desenvolvido na linguagem SQL.

Utilizou-se de três aplicações práticas para validação do algoritmo. Na primeira aplicação foram utilizados simulados. Na segunda aplicação o algoritmo foi aplicado em um conjunto de dados reais, os estratos gerados sobre cada um dos métodos de alocação e o também o GRD podem ser vistos nos anexos. Para este conjunto de dados observamos que a o método de alocação uniforme apresentou uma seqüência de estratos mais coerentes. Para os demais métodos, baseados na probabilidade, não houve uma redução tão acentuada verificado pelo EPA. Na aplicação 3, parte do algoritmo foi incorporado dentro de um SGBD, demonstrando-se com isso e flexibilidade do algoritmo também aplicado em uma base de dados reais.

Algumas observações devem ser feitas para o uso desta técnica:

- Há a necessidade de uma pesquisa anterior ou amostra piloto;
- O conjunto de dados deve conter características relevantes à característica de interesse;
- Há na maioria das vezes necessidade da limpeza e preparação dos dados;
- O processo de categorização das características contínuas utilizado é baseado na divisão dos valores contínuos em uma quantidade fixa de categorias, estabelecida pelo pesquisador;
- O algoritmo pode ser facilmente inserido nos SGBDs através da programação de visões, funções ou funções de armazenamento.;
- Relatórios e consultas que possuem a mesma variável de interesse podem compartilhar a mesma expressão de consulta.
- O algoritmo pode ser utilizado no processo de seleção de elementos para as rotinas de mineração de dados.
-

Sugestões que possam ser pesquisadas para melhorar o processo de identificação de estratos:

- Estudar e implementar outros métodos para categorização de valores contínuos;
- Implementar a categorização ordinal;
- Implementar o tratamento para alocação ótima com custo variável;
- Possibilitar a definição do processo de categorização de valores contínuos para cada característica;
- Estender o funcionamento dos procedimentos para utilização de várias tabelas na instrução de seleção;
- Estudar a validade de se implementar um dicionário no banco de dados para armazenar as informações sobre as expressões de consulta;
- Melhorar o procedimento para utilizar atributos binários ou dicotômicos na geração das expressões de consulta;
- Estudar estratégias para detecção e geração de índices para agilizar ainda mais o processo de consulta;
- Estudar a aplicação do GRD na analise discriminante.

REFERÊNCIAS

ABBEY, Michael; COREY, Mike e ABRAMSON, Ian. **Oracle9i – Guia Introdutório – Aprenda os fundamentos do Oracle 9i.** Editora Campus, Rio de Janeiro – RJ, 2002.

AURÉLIO, Marco; VELLASCO, Marley; LOPES, Carlos Henrique. **Descoberta de conhecimento e mineração de dados.** Pontifícia Universidade Católica, Laboratório de Inteligência Computacional Aplicada, 1999.

BARBARÁ, Daniel; CHEN, Ping. **Using self-similarity to cluster large data sets**. *Data Mining and Knowledge Discovery*, v. 7, n. 2, p. 123-152, Apr. 2003.

BARNETT, V.; **Elements of Sampling Theory.** The English University Press Ltd. St. Paul´s House Warnick. Lane, London, 1974.

BOLFARINE, H., BUSSAB, W. O. **Elementos de Amostragem**. São Paulo, Editora Edgard Blücher, 2005.

CHEN, M. S.; HAN, J.;YU P. S. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 1996.

CHENG, C et al. **Entropy-based Subspace Clustering for Mining Numerical Data**: *In* Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)), San Diego, 1999.

COCHRAN, W. G. **Sampling Techniques**. New York: John Wiley and Sons, 1977.

COOD, E.F. (1970), "A Relational Model of Data for a Large Shared Data Banks". Communications ACM, Vol. 13, No. 06, pp 377-387.

DATE, C. J, **Introdução a Sistemas de Bancos de Dados**, Campos, 2003.

DINIZ, Carlos Alberto R.; LOUZADA NETO, Francisco. **Data mining: uma introdução**. São Paulo: Associação Brasileira de Estatística, 2000.

FANDERUFF, Damaris. **Oracle 8i - Utilizando SQL*Plus e PL/SQL.** São Paulo, Makron Books, 1ª Edição, 2000.

FAYYAD, U.M.; **Advances in Knowledge and Data Mining: Towards a Unifying Framework.** Second International Conference on KD & MD., Portland, Oregon, 1996.

FREITAS, Alex A.; LAVINGTON, Simon H. **Mining very large databases with parallel processing**. Kluwer Academic Publishers. 1998.

GARRISON, R. H. e NOREEN, E. W. **Contabilidade Gerencial**. Tradução José Luiz Paravato. 9 ed. Rio de Janeiro: LTC, 2001.

HAN, Jiawei; KAMBER, Micheline **Data mining – Concepts and Techniques.** San Francisco: Morgan Kaufmann Publisher, 2000.

HANSEN, Morris H; HURWITZ, William N & MADOW, William G. **Sample survey methods and theory.** Vol. I. John Wiley & sons, Inc. 1966.

HEDAYAT A.S.; SINHA B.K.; **Design and inference in finite population**; Wiley and Sons, 1991.

INMON, William H**. Como construir o Data Warehouse.** Rio de Janeiro: Campus, 1997.

JACKSON, Joyce. Data mining: a conceptual overview. *Communications of the Association for Information Systems*. v. 8, p. 267-296, Mar. 2002.

JOHNSON, Norman L. e SMITH, Harry. **New developments in survey sampling**, John Wiley & Sons, 1969.

KORTH, Henry F. e SILBERSCHATZ, Abraham**. Sistema de Banco de Dados. 3 ed.** São Paulo: Makron Books, 2001.

NEYMAN J., (1934), **On two different aspects of the representative method: The method of stratified sampling and the method of purposive selection.** J. Royal Stat. Soc. B. 97, 558–606.

QUILAN, J. R. **Induction of decision trees**. Machine Learning, Dordrecht, v. 1, p. 81-106, 1986.

RAMALHO, José Antônio. **Oracle8i.** São Paulo: Editora Berkeley Brasil, 1999.

REZENDE, Solange Oliveira et al. **Mineração de dados.** In: REZENDE, Solange Oliveira (Org.). *Sistemas inteligentes*: fundamentos e aplicações. São Paulo: Malone, 2003, p. 307-333.

SÄRNDAL, C. E, Swensson, B. and Wretman, J., **Model Assisted Survey Sampling**. Springer-Verlag, 1992.

SCHERER, Douglas; GAYNOR William Jr.; VALENTINSEN, Arlene e CURSETJEE, Xerxes. **Oracle 8i: Dicas e Técnicas.** Rio de Janeiro – RJ, Editora Campus, 2000.

SNEATH, Peter H.; SOKAL, Robert R. **Numerical taxonomy: the principles and practice of numerical classification.** San Francisco: W. H. Freeman, 1973.

STAIR, Ralph M., **Principles of Information Systems – Managerial Approach**, Boyd & Fraser, 1996.

SHANNON, C.E. **A Mathematical Theory of Communication**, Bell Syst. Tech. J., 27.

SILBERCHATZ, Abraham, KORTH, Henry F. e SUDARSHAN, S., **Database System Concepts**. McGraw-Hill Companies, 1999.

PESSOA, D. G. C.; SILVA, P. L. N.. Análise de dados amostrais complexos. In: 13o Simpósio Nacional de Probabilidade e Estatística; 1998 jul 27-31; Caxambu (MG). Caxambu: ABE; 1998.

YE, Nong. **Data mining.** 3 ed New Jersey: Lawrence Erlbaum Associates, 2003.

THOMPSON, Steven K. **Sampling**. Wiley & Sons, New York, 1992.

ANEXO 1

Amostra parcial dos dados do Exemplo 2 do Capítulo 4.

	QCProd1	VVProd1	QCProd2	VVProd2	QCProd3	VVProd3	QCProd4	VVProd4	QCProd5	VVProd5	QCProd6	VVProd6	QCProd7	VVProd7	QCProd8	VVProd8	QCProd9	VVProd9	QCProd10	VVProd10	RegiaoCliente	Sexo	GrauInstrucao	EstadoCivil
0010000010015	9	20	0	0	0	0	2	7.01	7	10	0	0	0	0	0.65	2.98	0	0	0	0	1	Feminino	SupeRior Incompleto	Solteiro(a)
0010001000015	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	999	MasCulino	Superior Completo	Casado(a)
0010100000015	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Masculino	Segundo Grau	Casado(a)
0010100010014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Feminino	Segundo Grau	Solteiro(a)
0010100020013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	Masculino	Superior Completo	Outros
0010100030012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Feminino	SegUndo Grau	Casado(a)
0010100030029	0	0	0	0	0	0	0	0	4	9	0	0	0	0	0	0	0	0	0	0	1	Feminino	SegUndo Grau	Casado(a)
0010100040011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Feminino	Segundo Grau	Solteiro(a)
0010100050010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Masculino	Superior Incompleto	Solteiro(a)
0010100060019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Feminino	Segundo Grau	Outros
0010100070018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	999	MAsculino	Superior Completo	Solteiro(a)
0010100080017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Feminino	Superior Completo	Divorciado(a)
0010100080024	1	3	0	0	0	0	0	0	1	2	0.638	2.29	0	0	0	0	0	0	0	0	1	Feminino	Superior Completo	Divorciado(a)
0010100090016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Feminino	Superior Completo	Solteiro(a)
0010100100012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Feminino	Segundo Grau	Solteiro(a)
0010100110011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	MasculinO	Superior Completo	Casado(a)
0010100120010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	Masculino	Superior Incompleto	Solteiro(a)
0010100150017	1	2	0	0	0	0	0	0	0	0	0.775	3.86	0	0	0	0	0	0	0	0	1	Feminino	Superior Completo	Casado(a)
0010100160016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Masculino	Superior Incompleto	Solteiro(a)

ANEXO 2

Resultados do Exemplo 2 do Capítulo 4.

UNIFORME					
Estratos		**Qtde**	**Média**	**Variâncias**	
1	(EstadoCivil = 'Casado(a)') E (Sexo = 'MasCulino') E (GrauInstrucao = 'Superior Completo')	1303	0.00767	0.00915	15534.95235
2	(EstadoCivil = 'Casado(a)') E (Sexo = 'MasCulino') E (GrauInstrucao <> 'Superior Completo')	1110	0.00720	0.01977	24354.9207
3	(EstadoCivil = 'Casado(a)') E (Sexo <> 'MasCulino') E (GrauInstrucao = 'Superior Completo')	1674	0.00777	0.01248	34985
4	(EstadoCivil = 'Casado(a)') E (Sexo <> 'MasCulino') E (GrauInstrucao <> 'Superior Completo')	1646	0.00304	0.00303	8205
5	(EstadoCivil <> 'Casado(a)') E (RegiaoCliente = '1') E (GrauInstrucao = 'Superior Completo')	1895	0.00580	0.00999	35884
6	(EstadoCivil <> 'Casado(a)') E (RegiaoCliente = '1') E (GrauInstrucao <> 'Superior Completo')	4359	0.00206	0.00206	39231
9	(EstadoCivil <> 'Casado(a)') E (RegiaoCliente <> '1') E (Sexo = 'Feminino')	2799	0.00250	0.00535	41936
10	(EstadoCivil <> 'Casado(a)') E (RegiaoCliente <> '1') E (Sexo <> 'Feminino')	2379	0.00336	0.00419	23726
		17165		**Variâncias**	**0.0030390847**

ÓTIMA					
Estratos		**Qtde**	**Média**	**Variâncias**	
1	(Sexo = '') and (RegiaoCliente = '999') and (EstadoCivil = '')	296	0.00675	0.01347	34.35003162
2	(Sexo = '') and (RegiaoCliente = '999') and (EstadoCivil <> '')	242	0.00000	0.00000	0
3	and (Sexo = '') and (RegiaoCliente <> '999')	1613	0.00000	0.00000	0
4	(Sexo <> '') and (QCProd5 >= 35.1808) and (QCProd5 < 70.3616)	122	0.04098	0.03930	24.18555767
5	and (Sexo <> '') and (NOT ((QCProd5 >= 35.1808 and QCProd5 < 70.3616))) and (RegiaoCliente = '999')	688	0.00000	0.00000	0
6	(Sexo <> '') and (NOT ((QCProd5 >= 35.1808 and QCProd5 < 70.3616))) and (RegiaoCliente <> '999') and (RegiaoCliente = '4')	532	0.00000	0.00000	0
7	and (Sexo <> '') and (NOT ((QCProd5 >= 35.1808 and QCProd5 < 70.3616))) and (RegiaoCliente <> '999') and (RegiaoCliente <> '4') and (GrauInstrucao = 'Superior Completo')	4685	0.00704	0.01084	487.6902727
8	and (Sexo <> '') and (NOT ((QCProd5 >= 35.1808 and QCProd5 < 70.3616))) and (RegiaoCliente <> '999') and (RegiaoCliente <> '4') and (GrauInstrucao <> 'Superior Completo')	8987	0.00345	0.00588	689.1335674
		17165		**Variâncias**	**0.005179634**

PROPORCIONAL					
Estratos		**Qtde**	**Média**	**Variâncias**	
1	(QCProd5 >= 35.1808) E (QCProd5 < 70.3616)	126	0.03968	0.03810	4.8006
2	(NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao = 'Superior Completo') E (QCProd4 >= 0) E (QCProd4 < 0.9) E (RegiaoCliente = '5')	160	0.01875	0.03089	4.9424
3	(NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao = 'Superior Completo') E (QCProd4 >= 0) E (QCProd4 < 0.9) E (RegiaoCliente <> '5') E (RegiaoCliente = '1')	4186	0.00668	0.01047	43.810676
4	(NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao = 'Superior Completo') E (QCProd4 >= 0) E (QCProd4 < 0.9) E (RegiaoCliente <> '5') E (RegiaoCliente <> '1')	725	0.00000	0.00000	0
5	E (NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao = 'Superior Completo') E (NOT ((QCProd4 >= 0 E QCProd4 < 0.9)))	119	0.01680	0.01652	1.96588
6	(NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao <> 'Superior Completo') E (EstadoCivil = 'Outros') E (Sexo = 'Feminino')	300	0.00667	0.00662	1.986
7	(NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao <> 'Superior Completo') E (EstadoCivil = 'Outros') E (Sexo <> 'Feminino')	102	0.00980	0.00970	0.9894
8	(NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao <> 'Superior Completo') E (EstadoCivil <> 'Outros') E (GrauInstrucao = 'Primßrio')	237	0.00843	0.00836	1.98132
9	E (NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao <> 'Superior Completo') E (EstadoCivil <> 'Outros') E (GrauInstrucao <> 'Primßrio') E (GrauInstrucao = '')	3340	0.00089	0.00150	4.99664
10	E (NOT ((QCProd5 >= 35.1808 E QCProd5 < 70.3616))) E (GrauInstrucao <> 'Superior Completo') E (EstadoCivil <> 'Outros') E (GrauInstrucao <> 'Primßrio') E (GrauInstrucao <> '')	7870	0.00317	0.00596	46.9052
		17165		**Variâncias**	**0.006546934**

GRD							
Estratos		**Qtde**	**Média**	**Variâncias**	**Uniforme**	**Proporcional**	**Ótima**
1	(VVProd10 >= 0) E (VVProd10 < 7.616) E (QCProd5 >= 0) E (QCProd5 < 35.1808) E (QCProd1 >= 0) E (QCProd1 < 18.6) E (VVProd4 >= 0) E (VVProd4 < 3.321) E (RegiaoCliente = '1')	10406	0.00432	0.00718	777485.1225	74.71508	881.7511681
2	(VVProd10 >= 0) E (VVProd10 < 7.616) E (QCProd5 >= 0) E (QCProd5 < 35.1808) E (QCProd1 >= 0) E (QCProd1 < 18.6) E (VVProd4 >= 0) E (VVProd4 < 3.321) E (RegiaoCliente <> '1')	6110	0.00278	0.00474	176879.4898	28.94918	420.5704338
3	(VVProd10 >= 0) E (VVProd10 < 7.616) E (QCProd5 >= 0) E (QCProd5 < 35.1808) E (QCProd1 >= 0) E (QCProd1 < 18.6) E (NOT ((VVProd4 >= 0 E VVProd4 < 3.321))) E (GrauInstrucao = 'Superior Completo')	106	0.01886	0.01851	207.97836	1.96206	14.42145485
4	(VVProd10 >= 0) E (VVProd10 < 7.616) E (QCProd5 >= 0) E (QCProd5 < 35.1808) E (QCProd1 >= 0) E (QCProd1 < 18.6) E (NOT ((VVProd4 >= 0 E VVProd4 < 3.321))) E (GrauInstrucao = 'Superior Completo')	108	0.00925	0.00917	106.993872	0.990684	10.34378422
5	(VVProd10 >= 0) E (VVProd10 < 7.616) E (QCProd5 >= 0) E (QCProd5 < 35.1808) E (NOT ((QCProd1 >= 0 E QCProd1 < 18.6)))	150	0.00660	0.00662	148.95	0.993	12.20450736
6	(VVProd10 >= 0) E (VVProd10 < 7.616) E (NOT ((QCProd5 >= 0 E QCProd5 < 35.1808)))	174	0.02298	0.02246	679.99896	3.90804	26.07678968
7	(NOT ((VVProd10 >= 0 E VVProd10 < 7.616)))	111	0.00900	0.00892	109.90332	0.99012	10.48347843
		17165		**Variâncias**	**0.0097301192**	**0.0065545100**	**0.0064247404**